

Evaluating Administrative Records to Inform Measurement Error Properties of National Survey of College Graduates Estimates: Employment History and Firm Characteristics

by

Michaela Dillon
U.S. Census Bureau

CARRA Working Paper 18-11

September, 2021

This paper was written in 2018 under the former Center for Administrative Records Research and Applications program, but not published until 2021. CARRA has since been combined with the Center for Economic Studies, and the CARRA working paper series has been discontinued.

To obtain information about the series, see www.census.gov/ces or contact Christopher Goetz, Editor, Discussion Papers, U.S. Census Bureau, Center for Economic Studies 5K038E, 4600 Silver Hill Road, Washington, DC 20233, CES.Working.Papers@census.gov.

Abstract

Using administrative records in survey operations can potentially improve data accuracy and survey operations. In this study, we link administrative data on earnings from the Longitudinal Employer-Household Dynamics (LEHD) dataset to the National Survey of College Graduates (NSCG) to understand the alignment of this administrative records (AR) information with respondent collected data. The LEHD has ample coverage of the NSCG sample (over 90 percent) but has varying rates of agreement. The agreement rates across NSCG topics range from about 25 percent (for firm size) to about 75 percent (for employment status). Given these results, LEHD data has significant potential to enhance employment status and industry information within NSCG but may not be as useful for firm size and age.

Keyword: administrative records, college graduates, labor markets

JEL Classification: C83, E2, C8

* This paper is released to inform interested parties of research and to encourage discussion. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed, DRB approval number CBDRB-FY18-433.

EXECUTIVE SUMMARY

Using administrative records in survey operations can potentially improve data accuracy and reduce respondent burden. In this study, we link administrative data on earnings from the Longitudinal Employer-Household Dynamics (LEHD) dataset to the National Survey of College Graduates (NSCG) to understand the alignment of this administrative records (AR) information with respondent collected data.

The LEHD program populates its database with state unemployment insurance (UI) and Quarterly Census of Employment and Wages (QCEW) data linked to other government, administrative and Census Bureau records. This information allows the tracking of both aggregate and individual level employment, earnings, and job flows over time. Established in the 1970s, the NSCG is a biennial survey that collects data on the college-educated population of the United States, highlighting the connection between educational attainment and subsequent labor market outcomes.

This research evaluates conceptual alignment, linkage, and agreement of employment history and employer characteristics data between the LEHD and NSCG. Upon linking the LEHD to the NSCG by Protected Identification Key (PIK), the data shows ample coverage of the NSCG sample, over 90 percent. Among linked cases, the agreement rates—determined by a difference between NSCG and AR values of five percent or less—across NSCG topics are as follows: 74.87 percent for employment status, 31.96 percent for basic annual salary, 49.78 percent for firm age, 25.15 to 37.01 percent for firm size, and 53.49 percent for (non)education industry employment.

Some limitations with the data affect the conceptual alignment and agreement results of this research. Specifically, there is a coverage issue with LEHD data resulting in the underrepresentation of marginal workers such as the very young or old, retirees, students, or self-employed. In some cases the aggregated nature of the LEHD data presented some timing and definition issues that may overstate results within the employment status and salary topics. There is also some concern over the reporting of firm- and establishment-level data for firm size in the AR data, which produce unexpected patterns of agreement with NSCG responses by level of specification. The firm-level survey values agree more frequently with AR establishment-level than with firm-level information, 37 versus 25 percent, respectively..

Given these results, LEHD data has significant potential to enhance employment status and industry information within NSCG. The LEHD salary data displays similar results to other studies on measurement error in earnings and could be used for imputation purposes, even though it includes some forms of compensation not covered by the NSCG. The firm age information from the LEHD does not provide much added benefit, given low agreement rates. Finally, LEHD's usefulness for firm size is limited due to unclear documentation regarding firm-versus establishment-level information. Further investigation of the LEHD data and alternative data sources like federal tax data can inform future improvements to the measurement of employment and employer information.

TABLE OF CONTENTS

I. INTRODUCTION	6
II. LITERATURE REVIEW	7
III. DATA	8
3.1 National Survey of College Graduates (NSCG):	8
3.2 Longitudinal Employer-Household Dynamics (LEHD) Program:	9
3.3 Limitations	9
IV. RESEARCH QUESTIONS:	10
V. METHODOLOGY	10
5.1 Conceptual Alignment.....	10
5.2 Linkage.....	10
5.3 Agreement	11
VI. DATA MANAGEMENT	11
VII. RESULTS	12
7.1 Conceptual Alignment.....	12
7.1.1 Employment status	12
7.1.2 Basic Annual Salary	12
7.1.3 Employer Characteristics: Firm size, age, and industry	14
7.2 Linkage.....	14
7.3 Agreement	16
7.3.1 Employment Status.....	16
7.3.2 Basic Annual Salary	17
7.3.3 Employer Characteristics.....	19
VIII. CONCLUSION.....	26
IX. REFERENCES	28
X. APPENDIX	30

LIST OF TABLES AND FIGURES

Figure 1: Kernel Density Estimates of LEHD and NSCG 2010 Salary	13
Table 1: NSCG-LEHD Linkage Rates Across Respondent Characteristics	15
Table 2: Agreement in Employment Status across Demographics.....	16

Figure 2: Average Difference between LEHD and NSCG Salary by Percentile of the LEHD Earnings Distribution	18
Table 3: Agreement in Salary across Demographics	18
Table 4: Agreement in Firm Age	20
Table 5: Agreement in Employment Size (Firm)	21
Table 6: Agreement in Employment Size (Establishment).....	22
Table 7: Agreement in Employer Type (Education vs. Non-Education).....	23
Table 8: Agreement in Employer Type across Educational Institutions	24
Table 9: Agreement in Employer Type across Industries.....	25
Table 10: Most Frequently Reported Occupations for Linked Records in Disagreement on Employment Status	30

I. INTRODUCTION

The National Survey of College Graduates (NSCG) is a longitudinal survey of the college-educated population living in the United States. Sponsored by the National Center for Science and Engineering Statistics (NCSES) within the National Science Foundation (NSF), the survey informs two congressionally-mandated reports, *Women, Minorities, and Persons with Disabilities in Science and Engineering* and *Science and Engineering Indicators*, on the composition and productivity of the nation's STEM workforce. Thus, NCSES, with the Census Bureau serving as the data collection contractor, administers the NSCG to collect information on the human capital investment decisions and labor market outcomes of highly educated workers. Over time, the survey tracks respondents' demographic characteristics, educational attainment, workplace training, job satisfaction, professional mobility, and income.

NCSES is interested in the use of administrative data sources to enhance and supplement NSCG information. Administrative records have the potential to address many goals, including informing on measurement error, supplementing respondent-collected data, and reducing data collection and processing costs. To that end, NCSES has requested that the Demographic Research Group in the Center for Economic Studies (CES) of the Census Bureau (formerly the Center for Administrative Records Research and Applications) evaluate the NSCG for the use of administrative records to supplement and/or replace items on its questionnaire.

One promising source of administrative data with potential value to the NSCG is the Census Bureau's Longitudinal Employer-Household Dynamic Program (LEHD). The LEHD consists of several files structured at both the individual and firm/establishment level of analysis. The LEHD program gathers and organizes employment, earnings, and firm-level information over time from various sources including state agencies, business surveys and censuses, and federal tax forms. These data are used to improve imputation methods and to generate synthetic data for testing purposes with respect to employment information. Generally, historic data is available for all states from year 2000 to the most recent data release. Overall, the LEHD accounts for over 1.5 billion jobs, held by 262 million people, across 21 million firms.¹

In this study, we will assess the coverage, agreement, and quality of available LEHD data to respondent-provided information in the 2010 NSCG, specifically focusing on measurement error in employment history and questions about employer characteristics. This research addresses several objectives in the assessment of LEHD data for potential NSCG item supplementation or replacement. First, this work will identify which NSCG employment and earnings items could be enhanced by the information available within LEHD. Second, it will measure the extent to which measurement error exists within the data as determined by (dis)agreement between linked LEHD values and the distribution of responses between the two datasets. Finally, we will determine if data quality varies by key employment and demographic characteristics.

¹ Full documentation of the LEHD program is available here: <ftp://ftp2.census.gov/ces/wp/2014/CES-WP-14-26.pdf>.

II. LITERATURE REVIEW

This research investigates the ability of administrative records to replace and/or enhance survey data on employment history, earnings, and employer information. Specifically, the analysis includes identifying appropriate administrative data, linking it to survey responses, and evaluating the extent of measurement error of employment data between the two sources. The benefits to using AR with survey data are numerous. Künn (2015) emphasizes the reliability of AR to researchers as a routinely collected and authoritative data source. AR potentially lower data collection costs, circumvent the need for multiple datasets, and increase research opportunities via enhanced datasets. Additionally, the use of AR has been studied as a method to increase sampling efficiency for certain subpopulations, validate survey data, supplement survey data for difficult to obtain information, and improve forecasting ability of program costs (Bowie and Kasprzyk, 1987). An especially important benefit of AR is its ability to address various types of error within survey data.

The use of AR is one strategy to minimize item nonresponse, the instance when a respondent does not answer certain questions on the survey. Reasons for this behavior include concerns for privacy, cognitive difficulty/lack of understanding of survey item, risk of program eligibility, off-the-books earnings, and stigma of certain responses (Kunn, 2015; Bollinger et al., 2015). Because of respondent selection into nonresponse status or even intentional misreporting, bias within the data is not necessarily random across respondent characteristics. For example, nonresponse rates take on a U-shape across the income distribution where respondents at either end tend to omit income (Lillard et al., 1986; Bollinger et al., 2015; Korinek, 2005). Nonresponse also varies by occupation, educational attainment, and data collection mode (Lillard et al., 1986).

Another form of error addressed by AR is measurement error, the difference between the survey value and its true value. Comparison of linked survey and AR values sharing conceptual alignment sheds light on the extent to which measurement error is evident in survey values. For instance, large disagreement in linked values signals increased measurement error. Like item nonresponse, measurement error varies by respondent characteristics. Bollinger and David (2001) find it is higher for respondents that do not respond to all waves of a panel survey. Additionally, measurement error within longitudinal data is positively autocorrelated (Bound and Krueger, 1991). Additionally, Kreuter et al. (2014) observe measurement error among “hard to recruit” respondents already reluctant to participate in the survey.

This research uses AR sourced from business surveys that may be subject to nonresponse and measurement error as well. Regarding nonresponse error, Tomaskovic-Devey et al. (1994) detail several reasons a business might opt out of a survey surrounding motivation, capacity, and authority to respond. For example, the business may be avoiding the divulgence of confidential financial information, deferring questions to headquarters, or the assigned respondent for the organization is simply unavailable. Their analysis finds statistically significant increases in the probability of responding to surveys among price- and safety-regulated industries, as well as publicly traded companies.

When a company decides to respond to a survey, there are several opportunities for mismanagement or inaccuracy to compromise the integrity of data during the completion of a survey (Bavdaz, 2007). For instance, the risk of measurement error increases with the number of respondents in the firm that fill out requested information. Sometimes, the respondent may be external to the company, such as a contractor from an accounting firm that may not have thorough knowledge of requested information beyond recorded data. The quality of internal business data can also be a compromising factor depending on how businesses prioritize the maintenance of up-to-date and accurate information and minimize record error (Fecso and Pafford, 1988). Larger firms and foreign-owned companies tend to upkeep the quality of their data, whereas other firms put forth just enough effort to meet mandated requirements.

Survey design and implementation also inadvertently contributes to error when completion is presented as a legal requirement. Legal requirement improves the issue of nonresponse but has little effect on insuring the accuracy of responses. Error also varies by data collection mode. For example, if the respondent to a mail-in survey is not particularly motivated to contact survey administrators for clarification, respondent error may occur. This case is particularly problematic for recurring surveys where error may persist over several periods of data collection due to lack of communication and/or correction from the survey staff (Bavdaz, 2006; Bavdaz, 2007). Additionally, changes over time such as reorganizations, mergers and acquisitions, and personnel changes increase the probability of measurement error within business surveys.

III. DATA

3.1 National Survey of College Graduates (NSCG):

The NSCG is a biennial survey sponsored by the National Center for Science and Engineering Statistics (NCSES) within the National Science Foundation, administered by the Census Bureau, and sampled from the American Community Survey (ACS). It implements a rotating panel design in which respondents answer questions about their employment status, earnings, and education up to four times over a period of about six years. One of the unique features of the NSCG is its collection of data on more subjective information such as motivating factors for the individual's human capital investments, change in career or employment status. Additionally, the information collected in this survey informs two congressionally mandated reports on the U.S. STEM labor force: *Women, Minorities, and Persons with Disabilities in Science and Engineering*, and *Science and Engineering Indicators*. Survey respondents are college graduates, living in the U.S., up to age 75. Public data is available for download from 1993 to the present.

For this study, we use 2010 NSCG restricted access data. This particular year of the survey was the first data release after switching to its current sample frame, the ACS. To maintain the continuity of the rotating panel design, 46,828 new observations from the 2009 ACS were added to the sample already including 30,360 return respondents sampled from the 2001-2008 panels of the National Survey of Recent College Graduates and the 2003 NSCG for a total of 77,188 observations.

3.2 Longitudinal Employer-Household Dynamics (LEHD) Program:

The LEHD dataset is a collection of standardized data files sourced from administrative records on local employment, businesses, and earnings. Specifically, the data comes from state-level unemployment insurance (UI) records, the Quarterly Census of Employment and Wages (QCEW), as well as other administrative records sources available within the Census Bureau. The files link together via unique person and firm/establishment identifiers, the protected identification key (PIK) and state employer identification number (SEIN), respectively. Linkage allows the tracking of both aggregate and individual level employment, earnings, and job flows over time. The availability of this information varies by state depending on the data use agreement between the state data owners and the Census Bureau. The LEHD has nearly universal coverage of all states from the year 2000 forward.

This study uses the Employment History Files (EHF) and Employer Characteristics Files (ECF) within the 2011 release of the LEHD. The EHF provide quarterly earnings information for every job held by a worker as far back as year 1990 in some states. Therefore, each observation is a person-employer-year record of earnings. The universe numbers over three trillion observations. LEHD generates a record within the EHF only if the employer paid positive, nonzero wages during any quarter that year. Therefore, there are no missing values for annual earnings, the sum of quarterly earnings from an employer in a particular year, in the EHF.

The ECF provide quarterly employer microdata at both the firm and establishment levels. Relevant information is compiled from the QCEW, state UI records, imputations from the North American Industry Classification System (NAICS) , and control totals derived from data from the Bureau of Labor Statistics. Additionally, firm characteristic data on firm age and size are derived from Business Register and Longitudinal Business Dataset information. As it reflects national-level information for a firm, this information is especially useful in aligning with information requested by the NSCG. Each record is a year-quarter observation for each firm and establishment present in the QCEW or UI files.

3.3 Limitations

The LEHD files are an important resource that provides researchers a detailed description of worker employment histories. These histories, however, only include jobs and earnings reported to government agencies. Therefore, there is a coverage issue with LEHD data among certain occupations with informal payment practices. The LEHD underreports employment activity for those Abraham et al. (2013) characterize as marginal workers such as the very young or old, retirees, students, or self-employed. Difficulty also arises when trying to link survey respondents that simply choose not to work, since only the employed are present in the LEHD.

Additionally, federal employees are underrepresented in the EHF files as the Office of Personnel Management (OPM) manages their records separately. Federal employee earnings data includes different elements which are slightly inconsistent with wage data from state UI records. Also, a number of federal agencies, mostly national defense and justice, do not contribute data to this framework, citing security issues.

Lastly, states may join or opt out from the LEHD program at any time. Therefore, depending on the status of the individual agreement states have with the Census Bureau, LEHD's coverage by state may fluctuate over time.

IV. RESEARCH QUESTIONS:

The research questions are as follows:

1. To what extent are the concepts measured by the NSCG questions *aligned* with the administrative record information?
2. How often do NSCG records *link* to appropriate administrative record data that can be used to replace or supplement survey responses?
3. How often do data from the administrative records source *agree* with the responses from NSCG respondents by major subpopulation characteristics?

V. METHODOLOGY

The research questions of the previous section correspond to three analytical objectives of this research – to assess conceptual alignment, linkage, and agreement of linked information between the NSCG and LEHD. This section presents supplemental information on the analysis used to produce the data in the results section.

5.1 Conceptual Alignment

For research question #1, evaluation of conceptual alignment involves verifying that the data collected within both data sources are as similar as possible. Conceptual alignment also includes manipulating LEHD data when necessary to provide the closest approximation to requested NSCG information. The process of selecting appropriate observations within the linked dataset will be described in the following data management section. The results section will discuss simple coding to align LEHD information corresponding to specific NSCG questions. Also, specifically for the salary question, the analysis includes comparison of the salary distributions for the survey and administrative records data via kernel density estimates and plots of the average percent difference in linked values over the LEHD salary distribution. These visualizations highlight irregularities between the data sources as well as any outliers that will help guide research efforts to any areas of concern.

5.2 Linkage

For research question #2, in order to link the NSCG to the LEHD, both datasets require assignment of the unique identifier for individuals, the Protected Identification Key (PIK). PIKs allow linkage of information for a particular person across various Census surveys and administrative records. PIKs assignment to datasets occurs via the Personal Identification Validation System (PVS), a probabilistic matching algorithm used for record linkage at the Census Bureau. This process uses personally identifiable information (PII) from the survey such as name, age, and address to search reference files containing all known transactions for an SSN. Once matching information is found in the reference files with a certain threshold of confidence, the unique PIK value replaces the PII found on the survey data file. See Wagner and Layne

(2014) for a detailed description of the PVS process. The linkage rate, based on unique PIKs, represents the proportion of the PIKed NSCG sample found in the LEHD database. The analysis also includes the calculation of the linkage rate across various respondent demographics.

5.3 Agreement

For question #3, the results present findings on agreement in response and on the behavior of measurement error. For employment status, the results table presents the distribution of (dis)agreement across linked cases. For primary salary, agreement is assessed by the percent difference between linked values. Specifically, reasonable agreement is designated as the LEHD value being within five percent of the NSCG value. This threshold follows from an in-depth analysis of annual earnings information using the same data. In that study, the modal frequency (24.33% of linked cases) of agreement was among differences of 2-5 percent. About 35 percent of linked cases were within five percent. For both employment status and salary, the analysis extends these agreement distributions across several demographic characteristics including gender, age, race/ethnicity, and citizenship. Disaggregating the overall distributions allows observation of potential correlations of measurement error with respondent characteristics as well as identification of outliers within the data. For the employer characteristics data, agreement is assessed via cross-tabulation tables. The results present overall agreement rates as well as a discussion of the impact of establishment vs. firm-level data as a source of measurement error.

VI. DATA MANAGEMENT

The 2010 NSCG contains 77,188 observations. After undergoing PVS processing to assign PIKs to this dataset, the PIK rate is 98.35% (76,000 unique PIKs). Survey data typically has high PIK rates (90-93%). Failure to receive a PIK often occurs when the quality of personally identifiable information is low, or individuals do not have connection to the formal economy or to federal or state programs. Thus, non-PIKed persons are likely to be a non-U.S. citizen, be unemployed, not have health insurance, live in poverty, be under 35 years of age, be a minority, or have less than a high school education (NORC, 2011). Since individuals of high socioeconomic status, such as college graduates, exhibit fewer of these characteristics, the higher than usual PIK rate for the 2010 NSCG is to be expected.

Next, the assigned PIKs were appended to the 2010 NSCG response file via the survey unique identifier, REFID. After identifying unique PIKs from the NSCG file, it was linked to the LEHD Employment History File (EHF), which is a person-job-level file. Therefore, each observation within the EHF represents the PIK-SEIN-SEINUNIT-YEAR (person-firm-establishment-year) combination, wherein quarterly earnings information is available for each job held by an individual.² Because an individual may hold more than one job at a time, this merge is a one-to-many match, resulting in a linked dataset of approximately 1.2 million person-job-year observations. Furthermore, expanded employer information from the Employer Characteristics File (ECF) were added to the dataset via the SEIN (firm) and SEINUNIT (establishment) identifiers.

² State Employer Identification Number (SEIN). SEINUNIT is an establishment-level identifier.

Many NSCG survey items ask for information on a primary employer, requiring a collapsing of the linked dataset, which contains multiple jobs per person. We select the employer which paid the plurality of total earnings in the year as the primary employer and discard other jobs.

Identifying that employer among linked LEHD data requires the following steps:

1. Sort the linked dataset by PIK, year, and ascending annual earnings.
2. From the linked dataset, drop records for years other than 2010. This deletion results in a dataset including all jobs for individuals employed in that year.
3. Collapse the dataset to unique PIKs by retaining the final observation for each PIK, which reports the employer paying the individual the highest amount in 2010. Delete all other repeated PIK observations.

VII. RESULTS

7.1 Conceptual Alignment

This section lists the 2010 NSCG questions of interests in the study, and describes the adjustments made to the LEHD data for comparison with NSCG responses. The questions fall into three groups within the analysis: employment status, salary, and firm characteristics.

7.1.1 Employment status

- NSCG item A1: *Were you working for pay or profit during the week of October 1, 2010?*

To investigate this question, the dataset was collapsed by unique PIKs within the NSCG sample, retaining LEHD information corresponding to 2010:Q3.³ If the respondent was linked to non-missing LEHD earnings information for that quarter, that person is considered employed (=1) within the LEHD data. This corresponds to the “Yes” response in NSCG. If LEHD does not provide any earnings data for that quarter, the respondent is considered not employed (=0) within the LEHD data. This corresponds to the “No” response in NSCG. The comparison of this employment status information is straightforward from this point; however, note that the LEHD reports aggregated quarterly information while the NSCG response is for a particular week. Therefore, LEHD likely overreports the fraction of the sample working during that week as observations taking on a value of 1 includes individuals that worked earlier or later in the quarter, but not during that specific week.

7.1.2 Basic Annual Salary

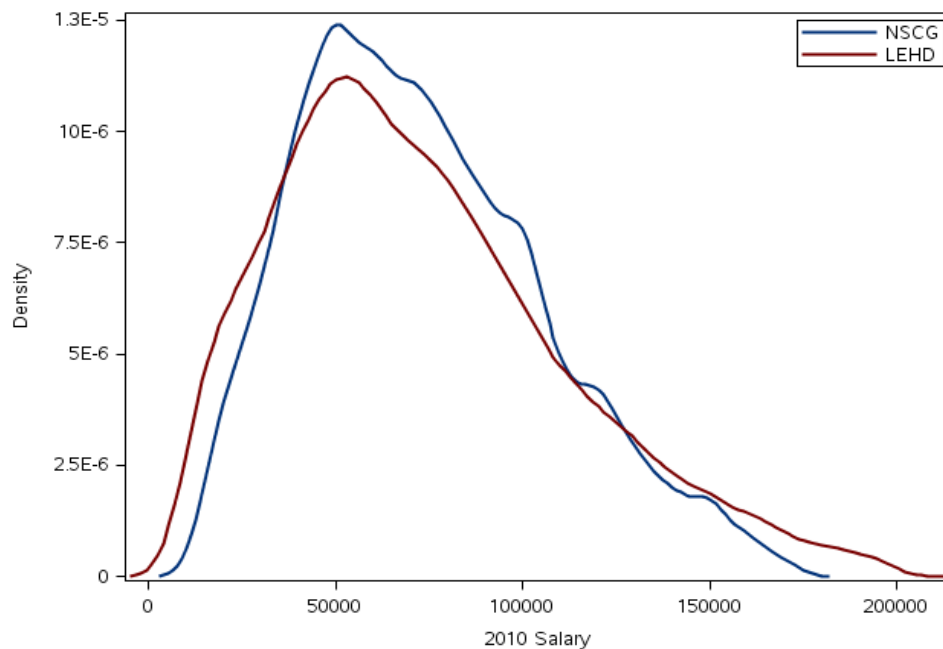
- A9: *Who was your principal employer during the week of October 1, 2010? If you had more than one job, report the one for which you worked the most hours that week. If your employer had more than one location report the location that employed you. If you worked for a contracting or consulting company, report the name of that company, not the client organization.*

³ Note: Oct. 1, 2010 is the start of the fourth quarter. However, since Oct. 1, 2010 occurred on a Friday, employment information for the third quarter was used in this analysis as it represents employment information for the majority of that week which is referenced in the NSCG question.

- A30: *As of the week of October 1, 2010, what was your basic annual salary on your principal job, before deductions? Do not include bonuses, overtime, or additional compensation for summertime teaching or research. If you are not salaried, please estimate your earned income, excluding business expenses.*⁴

These questions require the identification of a principal employer among a linked dataset allowing multiple jobs per person. NSCG describes the principal employer as the one for which the respondent worked the most hours. Since LEHD only contain hours information for a handful of states, this analysis assumes the principal employer is the one that paid the most earnings in 2010.⁵ Establishment identifiers are available within this dataset from which location information may be obtained via other LEHD files. The process for identifying the principal employer is described above in the data management section.

Figure 1: Kernel Density Estimates of LEHD and NSCG 2010 Salary



Source: 2010 NSCG linked to 2011 LEHD EHF files by PIK

Note that LEHD earnings information includes all payments to the employee during the quarter, including bonuses, overtime, and other compensation that NSCG requests be excluded from the response. Therefore, LEHD values may be relative to the NSCG question. Figure 1 shows the distribution of salary information between the two data sources. At the tails of the distribution, LEHD more frequently reports very low and very high earnings values. The literature on measurement error of earnings information supports this behavior among low-earning responses,

⁴ Subsequent research linking NSCG and LEHD data to evaluate the total annual earnings survey item is found in Dillon, M. (20XX) "Evaluating Administrative Records to Inform Measurement Error Properties of National Survey of College Graduates Estimates: An Analysis of the NSCG-LEHD Earnings Ratio."

⁵ Results are similar when using the max sum of 2009:Q4 through 2010:Q3 earnings to identify principal employment.

finding that low-income respondents tend to overreport their earnings relative to administrative data, and vice versa for high-income respondents.⁶ The inclusion of additional compensation beyond salary may be obscuring that result in the right tail (although there may also be misreporting of income for individuals whose compensation would be the conceptually the same in the NSCG and LEHD).

7.1.3 Employer Characteristics: Firm size, age, and industry

- A11: *Counting all locations where this employer operates, how many people work for your principal employer? Your best estimate is fine.*
- A12: *Did your principal employer come into being as a new business within the past 5 years?*
- A14: *Was your principal employer an educational institution?*
- A15: *Was the educational institution where you worked a...: grade school, junior/vocational college, four-year university, medical school, university research institute, other?*
- A10: *What was that employer's main business or industry—that is, what did that employer make or do? If your principal employer had more than one type of business, report the type of business primarily performed at the location where you worked.*

After identifying the primary employer as described above, two ECF files were merged via the SEIN-SEINUNIT identifier to obtain employer information from 2010 Q3. The first file contains information on national-level firm size and age. Indicators were coded to correspond to NSCG questionnaire responses.

The second file contained NAICS codes according to 2007 definitions. Two-digit level NAICS codes were used to create indicators for type of educational industry. NAICS definitions did not permit expansion of the LEHD data into the separate four-year college/university, medical school, and university-affiliated research institute responses available on the NSCG questionnaire. These categories are grouped into the Colleges, Universities, and Professional Schools category as defined by NAICS in the LEHD. In the agreement results, two-digit NAICS information (at the establishment-level) will be reported for all respondents.

7.2 Linkage

Linkage to the LEHD EHF files resulted in a linked dataset of approximately 1.2 million person-job-level observations. Collapsing the merged dataset to unique PIKs shows 93.95 percent of the PIKed NSCG sample as employed at some point in time covered within the LEHD.

As shown in Table 1, the linkage rate varies over some demographic characteristics. These results provide insight into whether AR provides sufficient coverage of certain groups of people. Looking at the first two columns, variation in the linkage rate occurs over age groups ranging from 86.50 percent for workers age 65-75 to 96.30 percent for workers in their twenties. Other respondent characteristics that display noticeable differences in linkage rates are ethnicity and

⁶ Bee, 2013; Roemer, 2002; Brummet et al., 2017

citizenship. Hispanic (87.08%) and non-U.S. citizen (89.35%) respondents link less frequently to LEHD than non-Hispanics (94.69%) and those with U.S. citizenship (94.32%).

Table 1: NSCG-LEHD Linkage Rates Across Respondent Characteristics

	Full Dataset		Selected Subset	
	Count	Linkage Rate	Count	Linkage Rate
Overall	76,000	93.95	52,500	95.60
Male	42,500	94.17	28,000	96.42
Female	33,000	93.68	24,500	94.65
21-29	8,700	96.30	6,600	96.58
30-39	17,500	96.88	14,500	96.91
40-49	17,500	95.29	13,500	95.91
50-64	24,000	92.60	18,000	93.98
65-75	8,200	86.50	-	-
Asian	12,000	94.61	8,700	95.44
Black	7,200	95.67	5,100	96.92
Multiple race	1,900	94.20	1,400	95.69
AIAN	400	95.68	300	96.72
NHPI	350	95.59	250	96.27
White	54,000	93.55	37,000	95.43
Hispanic	7,300	87.08	5,400	88.25
Non-Hispanic	68,500	94.69	47,500	96.44
U.S. citizen	70,500	94.32	48,000	96.10
Not a U.S. citizen	5,600	89.35	4,500	90.17

Source: 2010 NSCG and 2011 LEHD EHF files.

Note: In light of LEHD under-coverage of workers with low labor market attachment in the private sector, we show the linkage rate among a selected subset excluding respondents not of prime working age, employed part-time due to retiree or student status, a federal employee, or self-employed. Results rounded or suppressed (D) where necessary for disclosure avoidance.

Abraham et al. (2013) acknowledge under-coverage of certain workers by the LEHD exists. The last two columns of Table 1 recalculate the linkage rate within a subsample excluding “marginal” workers as a robustness check. Specifically, the reduced sample excludes workers age 65 or older, retirees, students, self-employed, and federal employees. Consequently, the overall linkage rate slightly improves to 95.60 percent. Likewise, the linkage rates across all the listed characteristics rise and maintain the relative relationships across demographic categories found

in the full sample. Note that there are particularly low linkage rates among older workers and non-citizens. Most workers begin to consider retirement in their sixties prompting an exit from the labor force, and non-citizens sometimes face barriers to entering the labor force due to visa restrictions. The following employment and salary results uncover similar underperformance in agreement among certain marginal worker groups.

7.3 Agreement

7.3.1 Employment Status

Overall, LEHD information agrees with NSCG responses on employment status during the reference week 75 percent of the time. The most frequent agreement between the two data sources occurs when both report that the individual is working (57.81%). This result is consistent across various demographics, with the exception of retirement-age workers. For these workers, the most frequent agreement is when both sources report the individual is not working (58.97%). Another result of interest is that the agreement rate increases with respondent age. Given the discussion on linkage, regarding under-coverage of certain marginal groups in the LEHD, this finding suggests LEHD information likely improves with worker tenure/experience. The agreement rate is further disaggregated among race, ethnicity, and gender; none of which display as much variation as the age groups. Regarding overall disagreement in employment status, when NSCG reports the respondent is working and LEHD does not, the respondent most frequently reports he is working 35 or more hours per week, is working less than 35 hours per week because he did not need/want to or due to family responsibilities. Conversely, when LEHD reports that the respondent is working and NSCG does not, the respondent most frequently reports he is looking for work, retired, or on layoff from a job.

Table 2: Agreement in Employment Status across Demographics

	Count	Agreement Rate	Employment Status			
			Agree		Disagree	
			Working	Not Working	NSCG Working ¹	LEHD Working
Overall	75,913	74.87	57.81	17.06	22.93	2.21
Gender						
Male	42,706	74.02	59.08	14.94	24.01	1.97
Female	33,207	75.95	56.17	19.78	21.54	2.51
Age						
21-29	8,707	67.83	57.33	10.50	29.53	2.64
30-39	17,299	73.06	63.86	9.20	25.12	1.82
40-49	17,430	75.85	67.61	8.23	22.73	1.43
50-64	24,241	75.71	58.58	17.14	21.91	2.37
65-75	8,236	81.53	22.56	58.97	14.75	3.72
Race						
White	53,885	74.89	57.36	17.52	23.05	2.06
Black	7,246	72.88	55.95	16.93	24.12	2.99
Asian	12,145	76.64	61.23	15.41	21.10	2.26

	Count	Agreement Rate	Employment Status			
			Agree		Disagree	
			Working	Not Working	NSCG Working ¹	LEHD Working
Pacific Islander	340	78.24	63.82	14.41	20.29	1.47
Native American	417	72.66	55.40	17.27	23.74	3.60
Multi	1,880	70.37	55.00	15.37	26.81	2.82
Ethnicity						
Hispanic	7,316	69.78	55.43	14.35	27.83	2.39
Not Hispanic	68,597	75.41	58.06	17.35	22.40	2.19
Citizenship						
U.S. Citizen	70,281	74.99	57.66	17.33	22.78	2.23
Not a citizen	5,632	73.35	59.62	13.73	24.75	1.90

Source: 2010 NSCG and 2011 LEHD files.

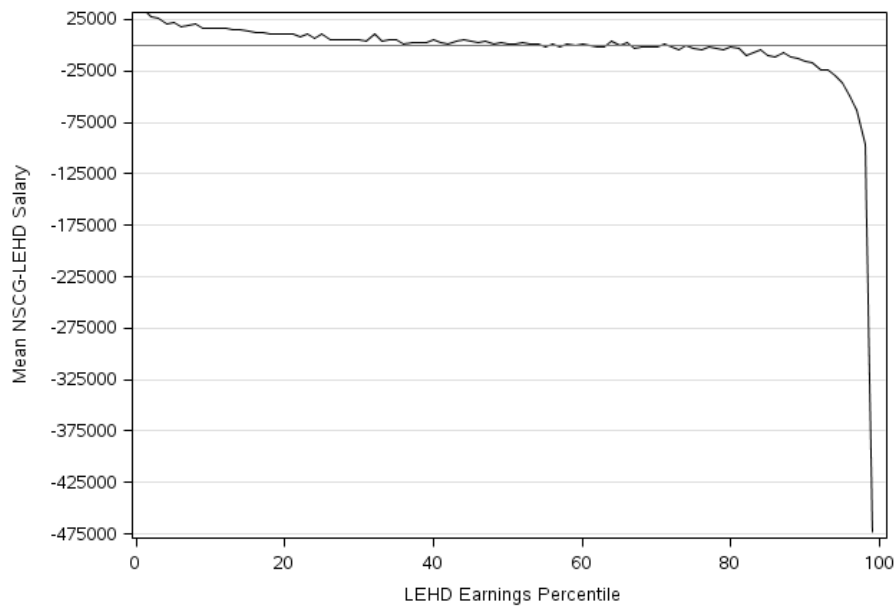
1: See Table 10 for a listing of the most frequently reported occupation among observations where the respondent indicates they are working and the LEHD does not.

7.3.2 Basic Annual Salary

For the salary analysis, the NSCG question requested earnings from a “principal” employer. As explained above, the survey and analytical definitions of the principal employer differ because the LEHD does not provide hourly information. Therefore, this research identifies that main employer as the one paying the most in 2010. After isolating the appropriate observations for principal employment, Figure 2 shows the relative relationship among linked values across the earnings distribution. The mean difference (measured NSCG value-LEHD value) between linked salary data is positive at the low end of the distribution and much of the middle portion, before becoming drastically negative among the highest percentiles. This same pattern is observed in a separate analysis on the NSCG annual earnings question and in a study linking IRS W-2 data to Consumer Expenditure Survey data (Dillon, 2018 and Brummet et al, 2018).⁷ It is a well-documented result that low-income respondents tend to overreport their earnings and upper-income respondents underreport relative to administrative records sources. That same result is reflected here.

⁷ Most of the analysis on basic annual salary aligns with the findings on NSCG annual earnings outlined in Dillon (2018). Refer to that study for more detailed analysis on how earnings information varies across the sample.

Figure 2: Average Difference between LEHD and NSCG Salary by Percentile of the LEHD Earnings Distribution



Source: 2010 NSCG and 2011 LEHD files.

Table 3 displays how agreement between linked values changes by respondent characteristics. Nearly a third of the sample has a linked LEHD value within five percent of the reported NSCG value. Also, the median percent difference (relative to the NSCG value) is 1.22% indicating over half of the sample has a percent difference greater than zero percent. This suggests NSCG has a tendency to underreport the salary value. Even though the fact that the LEHD value includes additional compensation that the NSCG question excludes, this result may be reliable since the same result is found in the analysis of annual earnings using an NSCG/LEHD earnings ratio.⁸

Table 3: Agreement in Salary across Demographics

	Count	5% Agreement	Median % Difference
Overall	43,470	31.96	1.22
Age			
21-29	5,150	28.39	-1.45
30-39	11,024	32.30	0.51
40-49	11,526	32.42	2.26
50-64	13,917	33.02	2.07
65-75	1,855	29.11	0.34
Race			
White	30,621	32.45	1.28
Black	4,041	33.23	0.53
Asian	7,327	29.53	1.54

⁸ 52% of the linked sample had an NSCG/LEHD annual earnings ratio greater than 1.

	Count	5% Agreement	Median % Difference
Pacific Islander	213	28.64	1.20
Native American	237	26.16	1.04
Multi	1,033	31.66	0.00
Citizenship			
U.S. Citizen	40,131	32.22	1.22
Not a citizen	3,341	28.85	1.21

Source: 2010 NSCG linked to 2011 LEHD EHF files by PIK.

Once again, there are notable deviations from average among the age groups. In particular, the youngest and oldest workers have lower rates of agreement between linked values. Volatile or sporadic employment driven by recent entry into or exit from the formal workforce may be driving these results previously discussed regarding the limitations of LEHD among marginal workers. Furthermore, workers in their twenties show a negative median percent difference indicating most of these respondents are linked to lower LEHD values. Knowing that earnings rise over the life cycle, this result aligns with what was seen in the kernel density estimates where there was an excess mass of LEHD values in the low end of the earnings distribution relative to NSCG values. It is also worthwhile to note that some minority populations have relatively low agreement rates in the data.

7.3.3 Employer Characteristics

The remaining results focus on employer characteristics. Agreement among the following topics depends upon the correct use of firm- or establishment-level data. LEHD contains both, and when correctly identified, can provide benefit to the NSCG via access to information beyond what is currently available to the survey.

Firm Age and Size

The firm age comes from an ECF file specifically designating nationwide, firm-level data. After creating indicators corresponding to less than five and greater than or equal to five years, the linked data shows agreement between the two data sources is virtually a coin flip. Summing the frequencies on the shaded diagonal of Table 4 and dividing by the linked sample size gives an overall agreement rate of 49.78 percent. Across responses, there is greater agreement among “new” firms, less than five years old.

Table 4: Agreement in Firm Age

NSCG Firm Age	LEHD Firm Age		
	Less than 5 years	5 or more years	Total
Less than 5 years	1,348 58.63	951 41.37	2,299
5 or more years	20,574 46.97	23,226 53.03	43,800
Unknown / Logical Skip	1,583 48.51	1,680 51.49	3,263
Total	23,505	25,857	49,362

Source: 2010 NSCG and 2011 LEHD ECF files.

The same ECF file also contains nationwide, firm-level employment size information, and by definition is conceptually well-aligned with the NSCG question which requests employment across all of the employer's locations. Table 5 displays the cross-tabulation of employment size responses for the linked sample. In spite of the LEHD's explicit documentation of firm-level information, row percentages in the "less than 10" column raise suspicions of significant measurement error. Agreement appears suitable among firms identified as smaller than 100 employees within the NSCG. Beyond 100 employees, LEHD greatly overreports firms as extremely small. With this data, overall agreement is around 25 percent.

Fortunately, LEHD has an alternative measure for employment size at the establishment-level within the EHF file. Agreement results using this information are reported in Table 6. The results along the shaded diagonal are more credible since they contain the largest row percentages. Agreement decreases as the size of the firm increases and exhibits a break in the downward trend when agreement goes from 50.74 percent for 100-499 employees to 27.24 percent for 500-999 employees. However, adding the row percentages to the immediate right or left of the diagonal value for 500-999 employees would smooth out the trend. Overall, agreement for this data is about 37 percent, and exhibits more logical results than in Table 5 which supposedly achieves better conceptual alignment. Whether the measurement error occurs from the respondent guessing their response or not understanding the context of the question, or from a structural issue within the administrative data is unclear.

Table 5: Agreement in Employment Size (Firm)

NSCG Number of Employees	LEHD Number of Employees								Total
	Less than 10	11-24	25-99	100-499	500-999	1,000-4,999	5,000- 24,999	25,000 or more	
Less than 10 employees	2,958 77.15	336 8.76	202 5.27	133 3.47	49 1.28	70 1.83	59 1.54	27 0.70	3,834
11-24	644 29.55	1,062 48.74	328 15.05	63 2.89	18 0.83	38 1.74	14 0.64	12 0.55	2,179
25-99	1,101 24.25	373 8.21	2,263 49.83	474 10.44	86 1.89	124 2.73	92 2.03	28 0.62	4,541
100-499	2,632 35.81	244 3.32	1,111 15.12	2,220 30.21	397 5.40	405 5.51	237 3.22	103 1.40	7,349
500-999	1,603 42.11	105 2.76	387 10.17	506 13.29	470 12.35	464 12.19	182 4.78	90 2.36	3,807
1,000-4,999	3,980 49.53	177 2.20	742 9.23	548 6.82	428 5.33	1,322 16.45	605 7.53	233 2.90	8,035
5,000-24,999	4,512 57.75	100 1.28	498 6.37	384 4.91	174 2.23	670 8.58	1,178 15.08	297 3.80	7,813
25,000 or more	5,937 69.51	69 0.81	316 3.70	277 3.24	113 1.32	263 3.08	625 7.32	941 11.02	8,541
Unknown / Logical Skip	1,682 51.55	229 7.02	447 13.70	325 9.96	101 3.10	211 6.47	180 5.52	88 2.70	3,263
Total	25,049	2,695	6,294	4,930	1,836	3,567	3,172	1,819	49,362

Source: 2010 NSCG and 2011 LEHD ECF T26 files.

Table 6: Agreement in Employment Size (Establishment)

NSCG Number of Employees	LEHD Number of Employees								Total
	Less than 10	11-24	25-99	100-499	500-999	1,000-4,999	5,000- 24,999	25,000 or more	
Less than 10 employees	2,680 69.90	289 7.54	171 4.46	214 5.58	87 2.27	199 5.19	145 3.78	49 1.28	3,834
11-24	450 20.65	1,094 50.21	275 12.62	105 4.82	46 2.11	102 4.68	70 3.21	37 1.70	2,179
25-99	354 7.80	443 9.76	2,490 54.83	601 13.23	109 2.40	290 6.39	182 4.01	72 1.59	4,541
100-499	450 6.12	203 2.76	981 13.35	3,729 50.74	847 11.53	684 9.31	306 4.16	149 2.03	7,349
500-999	162 4.26	89 2.34	283 7.43	905 23.77	1,056 27.74	953 25.03	267 7.01	92 2.42	3,807
1000-4999	304 3.78	139 1.73	482 6.00	1,192 14.84	913 11.36	3,450 42.94	1,252 15.58	303 3.77	8,035
5000-24999	246 3.15	93 1.19	400 5.12	1,042 13.34	642 8.22	2,195 28.09	2,676 34.25	519 6.64	7,813
25000 or more	241 2.82	117 1.37	368 4.31	992 11.61	812 9.51	2,395 28.04	2,496 29.22	1,120 13.11	8,541
Unknown / Logical Skip	468 14.34	192 5.88	399 12.23	636 19.49	316 9.68	678 20.78	418 12.81	156 4.78	3,263
Total	5,355	2,659	5,849	9,416	4,828	10,946	7,812	2,497	49,362

Source: 2010 NSCG and 2011 LEHD EHF files.

Industry of employment

NAICS information within the LEHD has the potential to expand the information available to the NSCG. The first two items in this section focus on educational institutions. However, since NAICS codes are available for over half of the linked cases, LEHD offers improved accuracy and heterogeneity of industry information. This is a great benefit to the survey and its end users. In Table 7, overall agreement on whether the respondent works in an educational institution is about 53 percent. The linked data more frequently agrees with NSCG responses among non-educational employers. Also, note the ability of LEHD to provide industry information for 57 percent of NSCG cases with an unknown response.

Table 7: Agreement in Employer Type (Education vs. Non-Education)

NSCG Employer Industry (NAICS 2-digit)	LEHD Employer Industry			
	Educational Institution	Non-Education	Unknown	Total
Educational Services	4,975 47.00	512 4.84	5,098 48.16	10,585
Non-Education	436 1.23	20,024 56.38	15,054 42.39	35,514
Unknown	422 12.93	1,438 44.07	1,403 43.00	3,263
Total	5,860	21,947	21,555	49,362

Source: 2010 NSCG and 2014 LEHD ECF files.

Note: NSCG industry value “Unknown” corresponds to records where the respondent did not answer the question due to logical skip for not working during the reference period. For LEHD industry, “Unknown” indicates a missing value.

Table 8 expands the information in Table 7 across several educational industry classifications. Even though some values are suppressed for privacy protections, linkage within these cells indicate NAICS information is available at a fine level of detail. The exception in this case is among four-year college, medical school, and university-affiliated research institute responses. Unfortunately, NAICS does not distinguish among those categories and must be compared to NSCG responses as one group. Among educational institutions, the linked data shows the greatest agreement among two-year colleges. Table 9 expands the information among non-educational institutions.

Table 8: Agreement in Employer Type across Educational Institutions

NSCG Educational Type	LEHD Educational Type						Total
	Preschool/ Grade School	2-Year College/ Comm/ Tech	4-Year College/ Univ Medical School Univ-Affiliated Research Institute	Other Education	Non-Educ. Institution	Unknown	
Preschool/Elementary/ Middle/Secondary	2,415	D	14	53	130	2,281	D
2-Year College/ Comm. College/ Tech. Inst.	D	521	10	29	19	437	D
4-Year College/University	12 0.41	47 1.61	1,319 45.25	22 0.75	168 5.76	1,347 46.21	2,915
Medical School	D	D	267	D	125	643	D
University-Affiliated Research Institute	D	D	239	D	65	385	D
Other Educational Institution	D	D	D	D	D	D	19
Non-Educational Institution	128 0.36	40 0.11	132 0.37	163 0.46	19,997 56.31	15,054 42.39	35,514
Unknown	226 6.93	40 1.23	129 3.95	27 0.83	1,438 44.07	1,403 43.00	3,263
Total	2,789	656	D	304	D	D	49,362

Source: 2010 NSCG and 2014 LEHD ECF files.

Note: NSCG industry value “Unknown” corresponds to records where the respondent did not answer the question due to logical skip for not working during the reference period. For LEHD industry, “Unknown” indicates a missing value.

Table 9: Agreement in Employer Type across Industries

LEHD Employer Industry (NAICS 2-digit)	NSCG Employer Industry			Total
	Educational Institution	Non-Education	Unknown / Logical Skip	
Educational Services	4,975 84.90	436 7.44	422 7.20	5,860
Non-Education	512 2.33	20,024 91.24	1,438 6.55	21,947
Agriculture, Forestry, Fishing and Hunting	7 4.58	133 86.93	13 8.5	153
Mining, Quarrying, and Oil and Gas Extraction	D	108	D	D
Utilities	D	160	D	D
Construction	11 1.89	519 89.18	52 8.93	582
Manufacturing	12 0.37	3,094 95.76	125 3.87	3,231
Wholesale Trade	D	1,355	94	D
Retail Trade	16 2.10	665 87.39	80 10.51	761
Transportation and Warehousing	D	265	21	D
Information	12 1.34	836 93.1	50 5.57	898
Finance and Insurance	D	1,017	49	D
Real Estate and Rental and Leasing	D	255	24	D
Professional, Scientific, and Technical Services	71 1.21	5,466 92.96	343 5.83	5,880
Management of Companies and Enterprises	D	318	14	D
Administrative and Support and Waste Management and Remediation Services	27 2.82	796 83	136 14.18	959
Health Care and Social Assistance	192 5.83	2,912 88.46	188 5.71	3,292

Table 9: Agreement in Employer Type across Industries

LEHD Employer Industry (NAICS 2-digit)	NSCG Employer Industry			Total
	Educational Institution	Non-Education	Unknown / Logical Skip	
Arts, Entertainment, and Recreation	12 4.43	222 81.92	37 13.65	271
Accommodation and Food Services	18 4.99	292 80.89	51 14.13	361
Other Services (except Public Administration)	88 12.01	563 76.81	82 11.19	733
Public Administration	26 2.33	1,021 91.32	71 6.35	1,118
Unknown	5,098 23.65	15,054 69.84	1,403 6.51	21,555
Total	10,585	35,514	3,263	49,362

Source: 2010 NSCG linked to 2014 LEHD ECF files. Values in the white rows are row percentages. To protect privacy, certain cells contain “D” indicating value suppression necessary for disclosure avoidance.

Using two-digit level NAICS codes, Table 9 lists all non-educational industries represented within the linked data. Due to space, the table is a transposed version of the previous cross-tabulation tables, such that the row percentages are the proportion of LEHD values in (dis)agreement with linked NSCG responses. The first two rows show NSCG agrees with LEHD’s assignment of (non)educational institutions at very high rates. Furthermore, nineteen additional industries are presented in the table, all of which is new information not currently provided to researchers that use NSCG data.

VIII. CONCLUSION

Summary of Results

This research evaluates conceptual alignment, coverage, and agreement of employment history and employer information between the National Survey of College Graduates (NSCG) and Longitudinal Employer-Household Dynamics (LEHD) data sources. These datasets were linked by PIK in order to achieve a person-job-level file from which to derive a comparable value for annual earnings across all jobs from the administrative data. The LEHD data provides very good coverage of the NSCG sample (93.95%). As for conceptual alignment, each topic fared well, with the exception of some timing and definition issues for salary information, and some concern over the use of firm- or establishment- level data for the firm size question. Aside from these issues, the data was easily manipulated to conform to NSCG responses.

Analysis of agreement between linked values found 74.87 percent of the linked sample agreed on employment status, and nearly a third (31.96%) of the linked LEHD salary data is within five percent of the NSCG value. There was some deviation from this overall result, particularly

among the youngest and oldest workers likely due to acknowledged limitations with the LEHD to capture workers with marginal formal employment.

The potential usefulness of LEHD data for employer characteristics information is mixed. On one hand, the firm age and size information is suspect until confirmation can be obtained on whether it is reporting firm- or establishment-level information. On the other hand, linkage of NAICS industry codes adds new information to the survey and is a great potential benefit to researchers that use NSCG data.

Recommendations for Future Work

The LEHD is an important and useful data source for employment information among formally employed individuals of prime working age. Essentially, the more stable the employment, the better the quality of the AR data. This assessment depends on the Census Bureau maintaining relationships with data providers, requiring the Bureau's commitment to implementing privacy protection policies and safeguarding against misuse in exchange for access to timely, quality administrative data on employment.

The above analysis shows LEHD data has significant potential to enhance employment status and industry information within NSCG. The LEHD salary data displays similar results to other studies on measurement error in earnings, even though it includes superfluous compensation. Therefore, it could be used for imputation purposes. The firm age information does not provide much added benefit beyond a 50-50 guess as to the appropriate age of the firm. As for firm size, it is not recommended to use that information at this time until the distinction between firm- and establishment-level information is verified.

Data limitations stem from underrepresentation of certain occupations and groups of people associated with marginal employment and/or informal payment practices. Also, in some cases, the level of detail requested in the NSCG is too fine for the administrative record to reflect. Specifically, this occurs when comparing responses for a reference week to quarterly information, as well as a subset of earnings information to aggregate compensation.

To overcome these limitations, it is necessary to further investigate the ability of LEHD data to confirm weekly employment and verify firm-level data is accurately assigned. Also, alternative data sources may be used to check the robustness of the results in this study. For example, federal tax information disaggregates some earnings information and may even provide new insight on how to identify principal employers. Also, summing the frequency of firm identifiers across the universe of W-2 data for a particular year could serve as an alternative measure for firm employment size.

IX. REFERENCES

- Abraham, Katharine G., John Haltiwanger, Kristin Sandusky, James R. Spletzer. 2013. "Exploring Differences in Employment between Household and Establishment Data". *Journal of Labor Economics* 31, no. 2: 129-s172.
- Bavdaz, Mojca. 2006. "The Response Process in Recurring Business Surveys". Proceedings of Q2006 European Conference on Quality in Survey Statistics.
<https://ec.europa.eu/eurostat/documents/64157/4374310/21-THE-RESPONSE-PROCESS-IN-RECURRING-BUSINESS-SURVEYS-SI-2006.pdf/1e30ecf5-2216-4dd9-82d7-3ed6b4f8d0ed>
- Bavdaz, Mojca, 2007. "Sources of Measurement Errors in Business Surveys". Presented at ICES-III, June 2007, Montreal, Quebec, Canada.
<https://ww2.amstat.org/meetings/ices/2007/proceedings/ICES2007-000032.PDF>
- Bollinger, Christopher R. and Martin H. David. 2001. "Estimation with Response Error and Nonresponse: Food-Stamp Participation in the SIPP". *Journal of Business and Economic Statistics*. Vol. 19 no. 2. 129-141.
- Bollinger, Christopher R., Barry T. Hirsch, Charles M. Hokayem, and James P. Ziliak. 2015. "Trouble in the Tails? Earnings Nonresponse and Response Bias across the Distribution". Working paper.
<http://economics.emory.edu/home/documents/Seminars%20Workshops/Seminar-2015-Hirsch.pdf>
- Bound, John and Alan B. Krueger. 1991. "The Extent of Measurement Error in Longitudinal Earnings Data: Do Two Wrongs Make a Right?" *Journal of Labor Economics*. Vol. 9 no. 1. 1-24.
- Bowie, Chester and Daniel Kasprzyk. 1987. "A Review of Administrative Records in the Survey of Income and Program Participation". SEHSD Working Paper no. 8721-43, U.S. Census Bureau, Washington, D.C.
- Brummet, Quentin, Denise Flanagan-Doyle, Joshua Mitchell, John Voorheis, Laura Erhard, and Brett McBride. 2018. "Investigating the Use of Administrative Records in the Consumer Expenditure Survey". CARRA Working Paper. U.S. Census Bureau, Washington, DC.
- Dillon, Michaela. 20XX. "Evaluating Administrative Records to Inform Measurement Error Properties of National Survey of College Graduates Estimates: An Analysis of the NSCG-LEHD Earnings Ratio". CARRA Technical Memorandum. U.S. Census Bureau, Washington, D.C.
- Fecso, Ron and Brad Pafford. 1988. "Response Errors in Establishment Surveys with an Example from an Agribusiness Survey".
https://ww2.amstat.org/sections/srms/Proceedings/papers/1988_057.pdf
- Künn, Steffen. 2015. "The Challenges of Linking Survey and Administrative Data". IZA World of Labor. <https://wol.iza.org/articles/challenges-of-linking-survey-and-administrative-data/long>
- Kreuter, Frauke, Gerrit Muller, and Mark Trappmann. 2014. "A Note on Mechanisms Leading to Lower Data Quality of Late or Reluctant Respondents". *Sociological Methods and Research*. Vol. 43 no. 3. 452-464.
- Lillard, Lee, James P. Smith, and Finis Welch. 1986. "What Do We Really Know about Wages? The Importance of Nonreporting and Census Imputation". *Journal of Political Economy*. Vol. 94 no. 3. 489-506.

- Tomaskovic-Devey, Donald, Jeffrey Leiter, and Shealy Thompson. 1994. "Organizational Survey Nonresponse". *Administrative Science Quarterly*. Vol. 39 no. 3. 439-457.
- Vilhuber, Lars and Devin McKinney. 2014. "LEHD Infrastructure Files in the Census RDC—Overview". CES Working Paper no. 14-26. U.S. Census Bureau, Washington, DC.
- Wagner, Deborah and Mary Layne. 2014. "The Person Identification Validation System (PVS): Applying the Center for Administrative Records Research and Applications' (CARRA) Record Linkage Software". CARRA Working Paper no. 2014-01. U.S. Census Bureau, Washington, D.C.

X. APPENDIX

Table 10: Most Frequently Reported Occupations for Linked Records in Disagreement on Employment Status

Principal job (minor group)	Reported occupations¹
Other non-S&E occupations	Lawyers, judges
Management-related occupations	Other management related occupations
Health-related occupations	Diagnosing/treating practitioners (dentists, optometrists, physicians, psychiatrists, podiatrists, surgeons, veterinarians)
Non-S&E managers	Top-level managers, executives, administrators (CEO/COO/CFO president, district manager, general manager legislator, chancellor, provost)
Sales and marketing occupations	Insurance, securities, real estate, and business services
Computer and information scientists	Software developers—applications and systems software; computer engineers—software
Biological and medical scientists	Biological scientists (botanists, ecologists, zoologists)
Social services and related occupations	Counselors (educational, vocational, mental health and substance abuse)
S&E managers	Engineering managers
S&E technicians and technologists	Electrical, electronic, industrial, and mechanical technicians

Source: 2010 NSCG and 2011 LEHD files.

1: Listing based on ranking of principal job categories as defined by variable N2OCPRNG in the 2010 NSCG in descending order of frequency. The top ten groups cover approximately two-thirds of linked records reporting disagreement in employment status (NSCG working-LEHD not working). Listed occupations (based on variable N2OCPR) represent the top 25 percent most frequently reported jobs for each group.